

LA-UR- 02-692

Approved for public release;
distribution is unlimited.

C.1

Title: SBE PRIMER: MULTIPLEXING
MINISEQUENCING-BASED GENOTYPING


Author(s): Lars Kaderali, University of Cologne (Germany)
Alina (nmi) Deshpande, B-2
Francisco Uribe-Romeo, B-2
Alexander Schliep, University of Cologne
David Torney, T-10 P. Scott White, B-1

Submitted to: Presentation at "The 10th International Conference on
Intelligent systems for Molecular Biology"
3-7 August 2002
Edmonton, Canada



Los Alamos

NATIONAL LABORATORY

Los Alamos National Laboratory, an affirmative action/equal opportunity employer, is operated by  University of California for the U.S. Department of Energy under contract W-7405-ENG-36. By acceptance of this article, the publisher recognizes that the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes. Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy. Los Alamos National Laboratory strongly supports academic freedom and a researcher's right to publish; as an institution, however, the Laboratory does not endorse the viewpoint of a publication or guarantee its technical correctness.



SBEprimer: Multiplexing Minisequencing-Based Genotyping

Lars Kaderali*¹, Alina Deshpande², Francisco J. Uribe-Romeo², Alexander Schliep¹, David C. Torney³ and P. Scott White²

¹Center for Applied Computer Sciences, University of Cologne, Weyertal 80, Cologne, 50931, Germany, ²Bioscience Division (B-1), Los Alamos National Laboratory, MS M-888, Los Alamos, NM, 87545, U.S.A. and ³Theoretical Division (T-10), Los Alamos National Laboratory, MS K-710, Los Alamos, NM, 87545, U.S.A.

ABSTRACT

Motivation: Single-nucleotide polymorphism (SNP) analysis is a powerful tool for mapping and diagnosing disease-related alleles. Most of the known genetic diseases are caused by point mutations, and a growing number of SNPs will be routinely analyzed to diagnose genetic disorders. Mutation analysis by polymerase mediated single-base primer extension (minisequencing) can be massively parallelized using for example DNA microchips or flow cytometry with microspheres as solid support. By adding a unique oligonucleotide tag to the 5' end of the minisequencing primer and attaching the complementary anti-tag to the array or bead surface, the assay can be "demultiplexed". However, such high-throughput scoring of SNPs requires a high level of primer multiplexing in order to analyze multiple loci in one assay, thus enabling inexpensive and fast polymorphism scoring. Primers can be chosen from either the plus or the minus strand, and primers used in the same experiment must not bind to one another. To genotype a given number of polymorphic sites, the question is which primer to use for each SNP, and which primers to group into the same experiment. Furthermore, a crosshybridization-free tag/anti-tag code is required in order to sort the extended primers to the corresponding microspheres or chip spots. These problems pose challenging algorithmic questions.

Results: We present a computer program to automate the design process for the assay. Oligonucleotide primers for the reaction are automatically selected by the software, a unique DNA tag/anti-tag system is generated, and the pairing of primers and DNA-Tags is automatically done in a way to avoid any crossreactivity. We report first results on a 45-plex genotyping assay, indicating that minisequencing can be adapted to be a powerful tool for high-throughput, massively parallel genotyping.

Contact:

kaderali@zpr.uni-koeln.de or scott.white@lanl.gov

(*) To whom correspondence should be addressed.

INTRODUCTION

Single nucleotide polymorphisms (SNPs) have been estimated to occur at a rate of about one every thousand nucleotides in the human genome (Cooper et al. (1985); Venter et al. (2001)). To date, more than 1.4 million SNPs have been identified, comprising a substantial proportion of all common human variation. Most of the known human genetic diseases are caused by point mutations, and tools to routinely analyze a growing number of SNPs will play a key role in medical diagnosis. Such tools will make it possible to perform association studies and linkage disequilibrium studies to identify genes that confer risk for common diseases (Schafer and Hawkins (1998)), and they will provide new insights into the history of human populations by allowing studies of human genetic diversity (Syvänen (1999)). Such applications could involve the simultaneous screening of thousands of SNPs, constituting a pressing need for robust, high-throughput and cost efficient SNP scoring methods.

The minisequencing approach to single nucleotide polymorphism analysis involves the annealing of an oligonucleotide primer directly adjacent to the mutation site, and polymerase mediated single-base extension using labeled dideoxynucleotide triphosphates (ddNTPs) (Syvänen (1999)). By combining this technique with the analytical power of flow cytometry using multiplexing microsphere arrays, Cai et al. (2000) demonstrate the applicability of the approach to the simultaneous analysis of multiple, potentially hundreds to thousands of sites (cf. also White and Torney (2001)).

Flow Cytometry measures fluorescence levels of particles at very high rates (hundreds to thousands of particles per second), and multiple fluorescence and scatter signals can be detected simultaneously (Nolan and Sklar (1998)). The technology is thus ideal for SNP analysis. By attaching unique DNA tags to the 5' end of each minisequencing primer, and by covalently binding the complementary tags (anti-tags) to carboxylated multiplexing (color-coded) microspheres, each primer binds to one specific microsphere.

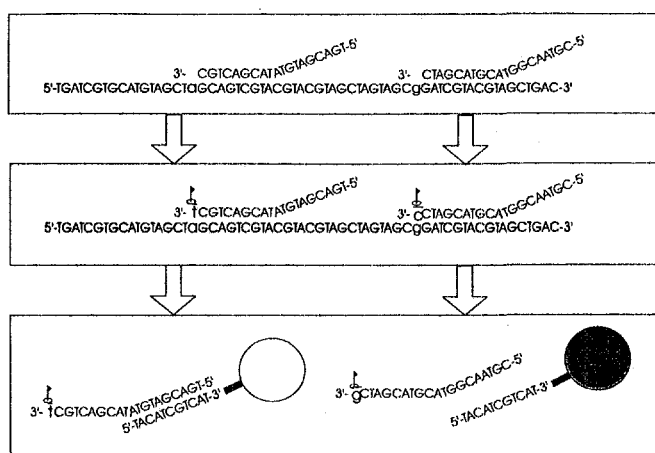


Fig. 1. Flow cytometry based minisequencing. Step one involves annealing of the primer-tag oligomers to the template strand, adjacent to the polymorphism sites. In step two, the polymerase extends the primer by fluorescently labeled ddNTPs, each base type bearing a different color. Step three involves dissociation of the primer and the template strand, and annealing between the tags and their microsphere bound antitags. Finally, microsphere color and base label reveal genotype.

The tagged microspheres are introduced to the reaction after the single base extension step, and the extended primer will anneal to its bound complement on the microspheres. By reading the fluorescent label of the single appended base and the microsphere color code using the flow cytometer, the mutations present at each site can be identified, and the identity of the site is known through the microsphere color. Figure 1 illustrates the experiment.

Cai et al. (2000) demonstrate that this technology permits multiplexed analysis of point mutations in genomic sequences. Flow cytometry has the great advantage of cheap assays and the wide availability of the required technology. Hirschhorn et al. (2000) have independently developed a similar assay using DNA oligonucleotide chips, with the antitags bound to the chip surface.

To permit multiplexed SNP scoring as described above, minisequencing primers must be appropriately chosen. Most importantly, such primers must not false prime, i.e., bind to a different site on the template strand (thus resulting in the incorporation of an arbitrary nucleotide unrelated to the SNP); they must not form homo- or heterodimer; they should not form hairpins; but they must bind immediately adjacent to the 3' side of the mutation and extend in the polymerase reaction. Furthermore, a tag / anti-tag oligonucleotide code is required in order to "sort" the extended primers to the corresponding elements on the microarray respectively to the correct microspheres in the flow cytometry assay, and thus "demultiplex" the experiment. These tags must not show any crosshybridiza-

tion, they must be carefully chosen to avoid reactions with any of the minisequencing primers, and the pairing of primers and tags poses another challenging combinatorial problem, as, here as well, hairpins and crossreactivity over the joint between primer and tag must be taken into account.

A number of computer programs exist to aid in the primer design process for the polymerase chain reaction (Dopazo and Sobrino (1993); Rychlik and Rhoads (1989); Lucas et al. (1991); Rozen and Skaletsky (1998); Giegerich et al. (1996) and others). However, some different constraints are required for minisequencing primers, that are not considered by the available software. Only one primer is required, and can be chosen from either the plus or the minus strand, as opposed to primer pairs in the PCR case. That primer must bind immediately adjacent to the 3' end of the polymorphism for the polymerase to append the next base opposite the SNP. Multiplexing of the minisequencing reaction requires very careful design of the oligonucleotides, as crossreactions between different minisequencing primers will inadvertently cause false results. Furthermore, all primers must work under the same reaction conditions, most importantly, the same temperature T . It is thus essential to use a most accurate model to predict nucleic acid hybridization.

THERMODYNAMICS OF DNA MELTING

Forces between nucleic acids in DNA duplexes are essentially of two kinds: Base pairing and base stacking. Contributions to the total stabilizing energy from base pairing depend exclusively on base pair composition, whereas contributions from base stacking depend on the actual basepairs formed and the base sequence along the chain. The latter are due to London dispersion forces and hydrophobic effects, and have only short-range effects. By assuming that they affect only the immediate neighbors of a given basepair, one derives the nearest neighbor model (Breslauer et al. (1986), compare also Owczarzy et al. (1997)).

The nearest neighbor model predicts free energy of nucleic acid binding (ΔG) based on the standard thermodynamic equation

$$\Delta G = \Delta H - T\Delta S, \quad (1)$$

where ΔH and ΔS are the enthalpy and entropy of duplex formation, and T is the reaction temperature in degrees Kelvin. The model assumes two-state-transitions, i.e. the DNA is either in the double helical or in the random coil, denatured state.

It is then further assumed that ΔH and ΔS can be calculated by summing up the contributions from the individual basepairs, taking into account the identity of their immediate neighbors. Thus, enthalpy ΔH_{duplex} of

the duplex is derived as

$$\Delta H_{\text{duplex}} = \sum_{i,j,k,l} N_{ij/kl} \Delta H_{ij/kl}, \quad (2)$$

where $N_{ij/kl}$ is the number of times a particular nearest-neighbor doublet ij/kl , ($i, j, k, l \in \{A, C, G, T, -\}$) appears in the duplex, i.e., the number of times an i/k basepair is followed by a j/l basepair in the duplex. Note that “—” indicates a “gap” in the duplex, meaning that the opposing nucleotide remains unpaired and bulges out of the helix.

The duplex melting entropy is determined in an analogous manner:

$$\Delta S_{\text{duplex}} = \sum_{i,j,k,l} N_{ij/kl} \Delta S_{ij/kl}, \quad (3)$$

and hence the duplex melting free energy is given by

$$\Delta G_{\text{duplex}}(T) = \sum_{i,j,k,l} N_{ij/kl} (\Delta H_{ij/kl} - T \Delta S_{ij/kl}). \quad (4)$$

The parameters $\Delta H_{ij/kl}$ and $\Delta S_{ij/kl}$ are usually derived from UV-absorbance versus temperature profiles of synthetic oligonucleotides (Allawi and SantaLucia (1997, 1998a,b,c); Breslauer et al. (1986); Gotoh and Tagashira (1981); Peyret et al. (1999); Quartin and Wetmur (1989); SantaLucia Jr. et al. (1996); SantaLucia (1998); Sugimoto et al. (1996)). By fitting the measured curves to the model, parameters can be obtained that according to SantaLucia Jr. et al. (1996) on average fit ΔG within 4%. Note that additional parameters are available to account for different buffer conditions and concentration effects.

ALGORITHM

Free Energy Calculation

Given two arbitrary DNA single strands, we now tackle the question whether they will form a stable duplex at some given temperature. We use the nearest neighbor model as described in the previous section for the thermodynamic considerations. However, its application requires prior knowledge about which basepairs form in the annealing reaction. The situation is further complicated as the duplex may contain bulges, i.e., unpaired bases within the duplex that bulge out of the double helix, but which do not impair stable binding of the two strands (Ke and Wartell (1995); LeBlanc and Morden (1991); Turner (1992)).

The Smith-Waterman alignment algorithm (Smith and Waterman (1981)) is widely used in bioinformatics to identify subsequences common to two given sequences. It calculates a local alignment between the two sequences and returns the optimum alignment found, maximizing or

minimizing a score

$$\sum w(x_i, y_j), \quad (5)$$

where $w(x_i, y_j)$ is a weight function over nucleotide or amino acid pairs and the summation is over all pairs formed in the alignment. The algorithm returns the optimum alignment of subsequences, allowing both gaps and mismatches in the two sequences. The general idea of the Smith-Waterman alignment algorithm is to calculate alignments of prefixes, and extend those until the optimum alignment of the entire sequences has been found.

We use the Smith-Waterman alignment algorithm to determine the minimum free energy alignment of two DNA strands at a given, fixed temperature T , using minimization of equation (4) as the objective function instead of equation (5) as in the original algorithm. Note that some modifications to the original algorithm are required to account for the nearest neighbors, however, this is relatively straightforward and can be done by storing additional information in the dynamic programming table. Thus, given two single DNA sequences, the algorithm computes the most stable interaction the strands can form, and returns free energy change ΔG for the nucleation reaction. If $\Delta G < 0$, the model predicts favorable energetics for the nucleation reaction, and we assume that the DNA strands interact in the experiment. Note that T is fixed in the calculation and should be set to the temperature at which primer and template annealing is carried out in the experiment, or some lower temperature if an additional “margin of safety” is desired.

Minisequencing Primer Selection

The SBEprimer program has been implemented to automate the minisequencing primer design process. The program designs sets of mutually compatible primers that will minimize the number of experiments required to genotype a given number of polymorphic sites on some template sequence. The program proceeds through seven iterative steps:

1. Read the input file containing the different SNP locations (with unique identifiers) and the template sequence.
2. For each SNP, check if either of the plus-strand or minus-strand primers adjacent to the SNP will false prime. If so, remove the primer.
3. Generate a list of all primers adjacent to SNP from both the plus and minus strand, that fulfill the requirements that their length is within some given limits and their melting temperature is higher than the temperature T_a at which primer annealing is carried out in the single base extension polymerase reaction.

4. Check each primer candidate for hairpins and homodimer formation; remove any candidate forming hairpins or homodimer.
5. Calculate all interactions between different primer candidates.
6. Choose maximal sets of mutually compatible primer candidates, such that each polymorphism has exactly one primer in exactly one primer set.
7. Output primer sets.

We discuss these steps in more detail in the following.

False Priming Check (Step 2) As has been pointed out before, the minisequencing primer for one polymorphism site can be chosen from either the plus or the minus strand, but it must bind immediately adjacent to the SNP. We speak of the “plus” and “minus” primer in the following; note however that for now we make no assumptions regarding the length of the primer, i.e., we do not fix its 5’ end at the time being. By choosing the primer such that its 3’ end binds adjacent to the SNP, the next base added to the primer by the polymerase will oppose the polymorphism and hence allow its typing. However, if the primer binds to a different location on the template strand and the polymerase extends there, false signal is generated. It is thus a requirement to exclude any primer that will false prime.

It has been shown that the polymerase reaction requires the terminal 3’ end bases of the primer to form stable basepairs with the template (Sommer and Tautz (1989)). Our procedure to identify potential false priming sites uses a hashtable of all 4-mers, for each such 4-mer storing a list of the plus and minus primers that contain the reverse Watson-Crick complement of that 4-mer at their 3’ end (considering two primers for each polymorphic site, the plus and the minus primer, and leaving their 5’ end open for the time being). This hashtable is then used in a routine that screens the entire template sequence and its complement. For each 4-mer in the template sequence, it checks the list of primers contained in the corresponding list in the table, and checks for potential false priming. The program attempts to extend the 4-mer duplex, checking if a stable interaction (with negative free energy ΔG) is possible. This check can be done by either a variant of the free energy calculation algorithm above, or by a simple base-by-base extension that does not allow any gaps. In either case, the extension is aborted if either a negative free energy interaction has been found (and the primer is consequently removed from the candidate set), or a threshold “maximum primer length” has been reached.

Note that the program permits to set the temperature for the false priming check separately, thus enabling the user to define more or less stringent conditions depending on individual requirements.

Primer candidate evaluation (Steps 3 and 4) The program will then evaluate the remaining candidates further. We will now consider the primer length as well, which has been neglected so far. To do so, SBEprimer will generate all primers from both plus and minus strand of each SNP (provided they have not been excluded in Step 2) of length between two given parameters *minlen* and *maxlen*.

The program will then calculate the free energy change ΔG associated with the binding of each primer candidate with its Watson-Crick complement, i.e., with the “intended” hybridization reaction of primer and template, for a given temperature T_a at which annealing will be carried out. A primer candidate is discarded if it has positive ΔG , as this means the primer will not bind and hence the single base extension step can not be carried out.

Subsequently, SBEprimer checks for homodimer and hairpin formation. The homodimer check is done by simply calculating the minimum free energy alignment of a primer candidate with itself. Again, a primer candidate is discarded if $\Delta G < 0$ for the minimum free energy alignment. Note that the temperature T for this calculation can be chosen independently from other steps involving free energy calculations, if desired. This allows more or less stringent conditions to be used in the different steps, depending on individual preferences and requirements.

We have tried several different simple heuristics for the hairpin check. Our results indicate that, whenever one of these predicts a hairpin, the homodimer check will also show potential homodimer formation (data not shown). We have thus decided against a dedicated hairpin check module, but assume this to be covered by the homodimer module as well.

Primer multiplexing evaluation (Steps 5 to 7) Finally, the primers must be chosen for the multiplexing assay. Given a set \mathcal{S} of SNPs and a list C_i , $i \in \{1 \dots |\mathcal{S}|\}$ of primer candidates for each $s \in \mathcal{S}$, the task is to generate disjoint sets $P_1 \dots P_m$ of primers that will fulfill the following criteria:

- For each SNP i , exactly one primer from C_i must be in $\bigcup_{j=1}^m P_j$. This means that each SNP is genotyped by a primer in one set.
- Any two primers p_a and p_b from the same set P_j “work together”, i.e. no heterodimer-formation occurs between any two primers within the same set. Hence, the primers in one set P_j can be multiplexed.
- The number m of different sets is minimized. m corresponds to the number of experiments that have to be run separately in order to genotype all polymorphisms. Ideally, $m = 1$.

The multiplexing module involves the prediction of all pairwise interactions between any of the primer candidates

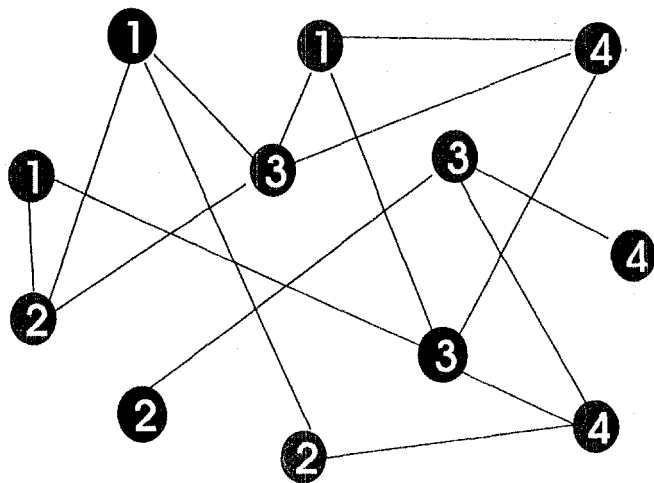


Fig. 2. Sample instance of the generalized graph coloring problem. The vertices correspond to primer candidates, two vertices being connected when the corresponding primers interact and genotype different polymorphisms. Two vertices with equal numbers type the same SNP. The dark black vertices constitute an optimum solution, as one vertex has been chosen for each polymorphism, and the corresponding primers do not interact. In this simple example, all chosen vertices can be used in one single experiment.

from distinct SNPs. Those interactions are calculated using the minimum free energy alignment algorithm, as described in the corresponding section above. Again, we assume that two given primers interact if $\Delta G < 0$, and will not interact otherwise.

We then use a graph theoretic model to solve the problem. Create an undirected Graph $G = (V, E)$ with vertex set V and edge set E as follows: For each primer candidate left after step 5, create one vertex $v \in V$. Furthermore, label each vertex with the identifier of the polymorphism genotyped by the corresponding primer. Finally, create an edge $(v_1, v_2) \in E$ between two vertices v_1 and v_2 , if the minimum free energy alignment for the corresponding primers shows negative ΔG , i.e., if the corresponding primers interact, and if they bear different SNP labels (and hence genotype different polymorphisms). Figure 2 illustrates the construction.

The multiplexing primer selection problem then transforms to a special version of the graph coloring problem. In the graph coloring problem, the task is to assign a color to each vertex, where no adjacent vertices (vertices v_1, v_2 connected by an edge $(v_1, v_2) \in E$) can have the same color, and the number of different colors used over the entire graph is to be minimized. In our case, not all vertices need to be assigned a color, but only one vertex from each group of vertices bearing the same label. Which one of them can be freely chosen. This construction ensures that

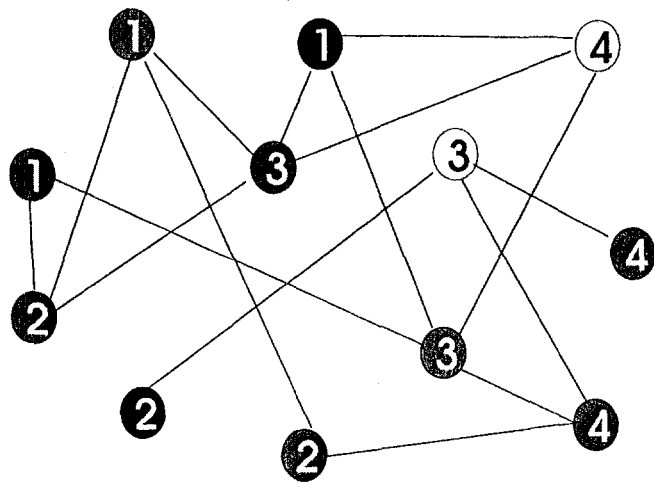


Fig. 3. This coloring also shows a feasible solution of the generalized graph coloring problem instance from Figure 2, requiring two distinct experiments. The primers corresponding to the dark black vertices can be used to genotype polymorphism sites 1 and 2, and the primers corresponding to the white vertices for SNP sites 3 and 4. Two experiments are required, as primer 1 and 4 as well as primer 2 and 3 will bind to one another and thus cannot be multiplexed in the same assay. The solution is thus suboptimal.

exactly one primer is used for each polymorphic site, and vertices respectively primers assigned the same color can be multiplexed together in the same experiment.

It is well known that the graph coloring problem is NP-complete (Garey and Johnson (1979)), i.e., it is widely believed that there exists no efficient algorithm to solve the problem exactly. Clearly, our version generalizes the problem further, as an algorithm to solve the extended version of the problem encountered in our case could easily be used to solve the original graph coloring problem, simply by assigning different labels to all vertices in the graph coloring problem instance. This would ensure that all vertices are assigned a color. On the other hand, given a solution for an instance of the generalized graph coloring problem, it is easy to check its feasibility. This proves that the generalized graph coloring problem is NP-complete as well, hence we can not expect to find a polynomial time algorithm for the problem.

We have thus decided to use a simple heuristic to find a feasible solution. In a first step, the algorithm splits the vertices in $|\mathcal{S}|$ distinct groups, according to their label. Then, all vertices v within each group are sorted according to their vertex degree

$$\deg(v) := |\{(x, y) \in E | v = x \text{ or } v = y\}|. \quad (6)$$

Only the vertex n with the lowest degree

$$\deg(n) = \min_{v \in V} \deg(v) \quad (7)$$

is kept in each group, all other vertices and the corresponding edges are removed from the graph. Hence, we now have an instance of the standard graph coloring problem, where each vertex has to be assigned a color. A simple greedy heuristic is finally used to color this residual graph.

Tag Generation

A second problem associated with “demultiplexing” the genotyping assay is the need for tag / anti-tag pairs to be used with the primers. The tags will be conjugated with the minisequencing primer, the resulting oligo having a dual function: While the primer part is required for the single base extension reaction, will the tag part sort the extended primer to its corresponding antitag on the bead or chip surface, thus enabling simple readout of the results.

For this to work, there must not be any crossreactivity between the tags and antitags. For a set T of tags and the set \bar{T} of the complementary antitags, the following conditions must hold:

1. $t_i \in T, \bar{t}_i \in \bar{T}$ should bind for all i ; i.e., each tag must bind to its corresponding antitag.
2. $t_i \in T, \bar{t}_j \in \bar{T}, i \neq j$ must not bind; i.e., a tag must not bind to a “foreign” antitag.
3. $t_i \in T, t_j \in T$ must not bind; i.e., no two elements in T should bind to one another (including the homodimer case).
4. $\bar{t}_i \in \bar{T}, \bar{t}_j \in \bar{T}$ must not bind; i.e., no two elements in \bar{T} should bind to one another (including the homodimer case).

The problem of designing DNA tag-antitag systems satisfying requirements (1) and (2) has been previously described. Frutos et al. (1997) use a coding theory approach with Hamming distance conditions to avoid crosshybridization. They design octamers with 50% G-C content, differing in at least 4 bases from each other. The approach followed by Brenner (1997) implies the construction of the largest possible λ -free code for a given λ . Morris et al. (1997) use De Bruijn sequences of order λ to obtain such λ -free codes. Ben-Dor et al. (2000) extend this approach to incorporate a simple thermodynamic model, employing the 2-4-rule: The melting temperature in degrees Celsius of a duplex is assumed to be equal to twice the number of A-T basepairs plus four times the number of G-C basepairs.

We felt these approaches had two shortcomings: First of all, interactions between different tags (and not involving any antitags) could also interfere with tag-antitag hybridization and thus a clear, strong signal. Secondly, and even more importantly, in order to obtain a high level of multiplexing one would clearly benefit from a more sophisticated thermodynamic model, taking into account, for example, effects of mismatches, bulges and dangling ends.

We have thus implemented the following greedy algorithm to generate such sets:

```

1 Start with empty set T
2 repeat (add tags)
3   repeat (generate sequence)
4     Generate a random sequence S
      and its complement S'
5   until S and S' form no homodimer
6   if both S and S' do not interact
      with any other sequence in T
7     add S, S' to T
8 until enough tags.
```

Where T is the set of tags and antitags generated so far. The generation of a new sequence S in line 4 is done base by base, drawing each base randomly and i.i.d. from the set $\{A, C, G, T\}$. New bases are added at the end of S until its melting temperature T_m reaches a predefined bound, say 60 degrees Celsius. This temperature is calculated as

$$T_m = \frac{\Delta H}{\Delta S + R \ln C_t}, \quad (8)$$

where ΔH and ΔS are enthalpy and entropy changes of duplex formation derived from the nearest neighbor model, R is the Boltzmann constant, and C_t is the total molar concentration of strands.

Again, we assume that a check for homodimer formation will also catch hairpins, and the interactions in line 6 are calculated using the free energy alignment algorithm described above.

Primer-Tag Pairing

One central idea of the assay is the independence of the minisequencing primers and the tag/antitag system. In principle, the same set of tags can be used for all experiments, and hence custom microsphere sets or DNA chips can be pre-fabricated and stored. One problem that needs to be considered is the possibility of interactions between primers and tags/antitags. This may force us to exclude certain tags for a specific assay. Note that we have so far checked for crosshybridization between the tags and antitags, but not considered the case where the primer binds to a tag or the combination between primer and tag leads to additional problems, as some undesired hybridization over the joint of the two may occur or primer foldback becomes a problem. Therefore, we also need to decide which tag to combine with which primer. The following problems need to be addressed:

- Binding of a tag-primer pair to an undesired antitag on the microspheres or chip surface, leading to wrong signal.

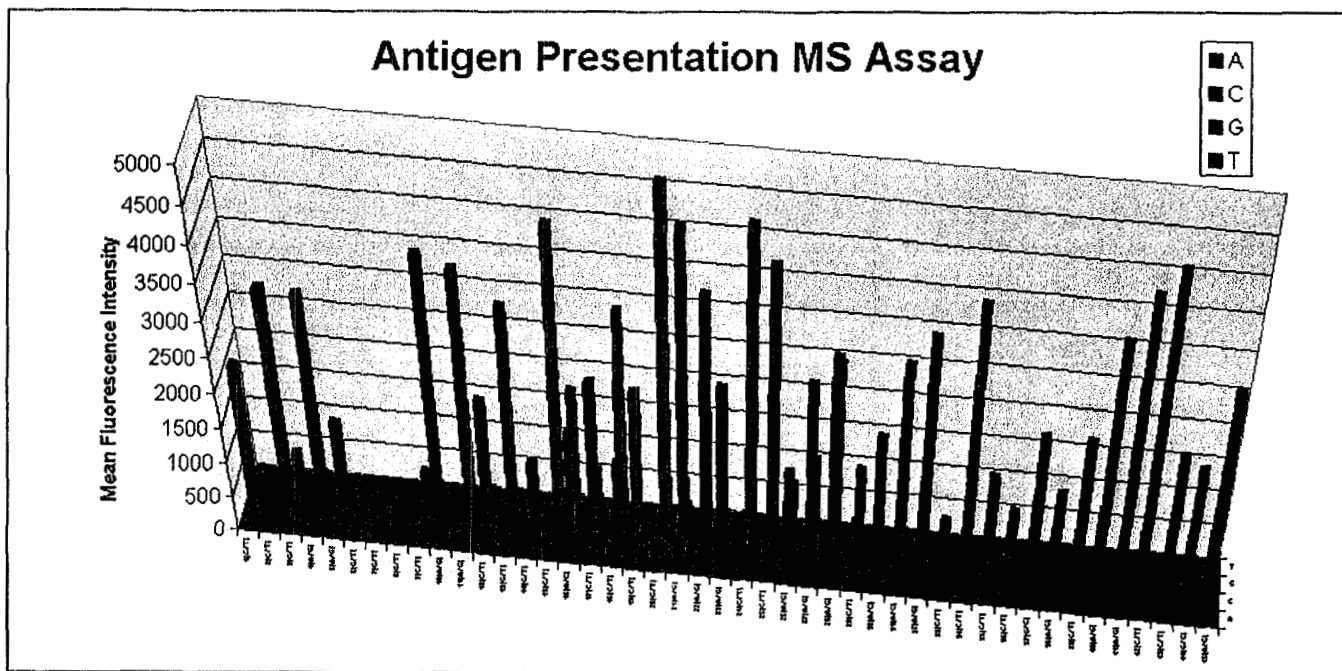


Fig. 4. 45-plex Genotyping trace. The plot shows absolute fluorescence levels for each of the 45 polymorphisms after subtraction of the background values obtained from a bead only control.

- Binding of different primer-tag pairs to one another, leading to false extension in the minisequencing reaction.
- Foldback of a primer-tag pair onto itself, causing wrong primer extension and lower signal on the antitags due to competitive reactions.

We have implemented a computer program as part of the SBEprimer package that will, given the set P of primers, a set T of tags with $|T| > |P|$ (i.e., we assume a larger number of tags to be given than primers), and the set \bar{T} of antitags, chose a subset of the tags and pair them with the primers, assuring that the above problems are avoided. The process is straightforward: Starting with a random pairing, the program will check for hairpins, dimer-formation and crossreactivity between the primer-tag oligonucleotide and the antitags. These checks are performed using the free energy alignment algorithm as in the previous sections. If any problem is encountered, the involved tag/antitag pair is discarded and replaced by a new pair from T and \bar{T} . This is iterated until either a feasible combination is found, or no more tags remain in T to exchange.

The program will then output the list of primer-tag pairs and antitags to use in a format that can be used to directly order or assemble the oligonucleotides required.

EXPERIMENTAL RESULTS

A set of 45 multiplexed minisequencing primers for scoring 45 SNPs in target genes in the human Major Histocompatibility Complex (MHC) was selected using the SBEprimer program. 45 tag-antitag pairs were automatically chosen for these primers from a set of 250 generated by the software, and a final list of oligonucleotides that had the sequence of the tag and the minisequencing primer was generated for synthesis.

The MHC complex is located on chromosome 6 in the human genome, and target regions containing the 45 SNP sites were amplified using the Polymerase Chain Reaction (PCR). 15 amplicons were generated using this technique. The amplicons were pooled together, and treated with Shrimp Alkaline Phosphatase, which removes the excess unused dNTPs (deoxy nucleotide triphosphates), and Exonuclease I, which removes single stranded primers used in the PCR reaction.

Single base extension was performed in a 10 μ l reaction that consisted of the pooled, clean template, thermosequenase (0.75 units), thermosequenase reaction buffer, biotinylated dideoxy nucleotide triphosphates (ddNTPs, 7.5 μ M), and 45 minisequencing primers (25 nM, each primer). A single reaction was performed for each of the four biotinylated ddNTPs. The cycling conditions used were an initial incubation at 94°C for 10 seconds, and annealing and extension of the single base at 60°C for 10

Bead	SNP ID	Poss. Alleles	A	C	G	T	% A	% C	% G	% T	Call	Conv. Sequencing
1	LMP2E3605U	C/T	3	2087	155	68	0,00	0,90	0,07	0,03	C	C
2	LMP7E1B702U	C/T	21,5	18	10,5	2733,5	0,01	0,01	0,00	0,98	T	T
3	LMP7E1B718L	G/T	67,5	39,5	620,5	23,5	0,09	0,05	0,83	0,03	G	G
4	LMP7E1B755U	A/G	73	-4	2923	84	0,02	(0,00)	0,95	0,03	G	G
5	T1E5907L	A/G	45	34	1145	18,5	0,04	0,03	0,92	0,01	G	G
6	T1E67153U	G/T	93	46	228	26,5	0,24	0,12	0,58	0,07	G	
7	T1E67225U	G/T	34	-13	243	32	0,11	(0,04)	0,82	0,11	G	
8	T1E67262L	G/T	408	18	67	13	0,81	0,04	0,13	0,03	A	
9	T1E7L	C/T	-4	5	528	-7	(0,01)	0,01	1,01	(0,01)	G	
10	T1E101063U	A/G	4073	14	107	18	0,97	0,00	0,03	0,00	A	A
11	T1E10990U	A/G	109	108	3449	23	0,03	0,03	0,93	0,01	G	G
12	T1E10983L	C/T	108,5	1899	375	78	0,04	0,77	0,15	0,03	C	C
13	T1E10908U	G/T	123	55	3025	89,5	0,04	0,02	0,92	0,03	G	G
14	T1E10906L	G/T	50	52,5	811,5	12,5	0,05	0,06	0,88	0,01	G	G
15	T2E1U	C/T	57,5	56,5	141,5	3934,5	0,01	0,01	0,03	0,94	T	T
16	T2E6L	A/C	166	1587	1933	105	0,04	0,42	0,51	0,03	G/C	C
17	T2E7524U	C/T	32	2323	786,5	58	0,01	0,73	0,25	0,02	C	C
18	T2E7534L	C/T	150	45	933,5	2872	0,04	0,01	0,23	0,72	T	T
19	T2E7584U	C/T	77	2248	265,5	32	0,03	0,86	0,10	0,01	C	C
20	T2E7782U	G/T	305,5	88,5	238,5	4598,5	0,06	0,02	0,05	0,88	T	T
21	T2E994L	A/G	308	52,5	4273	169	0,06	0,01	0,89	0,04	G	G
22	T2E9161U	A/G	92	185,5	3430,5	171	0,02	0,05	0,88	0,04	G	G
23	T2E9197U	A/G	2675,5	16,5	110	86	0,93	0,01	0,04	0,03	A	A
24	T2E9223L	C/T	335	606	354,5	4174	0,06	0,11	0,06	0,76	T	T
25	T2E9269U	C/T	83	56	59	3678	0,02	0,01	0,02	0,95	T	T
26	T2E9286L	A/G	1573,5	12	894	40	0,62	0,00	0,35	0,02	A/G	A/G
27	T2E10U	A/G	2842	89	1278	89	0,66	0,02	0,30	0,02	A/G	A/G
28	T2E111341U	A/G	3232,5	213	2670	38	0,53	0,03	0,43	0,01	A/G	A/G
29	T2E111275U	C/T	66	679	367	941	0,03	0,33	0,18	0,46	C/T	C/T
30	TPSN149L	A/C	38	1951	44	62	0,02	0,93	0,02	0,03	C	C
31	TPSN150L	A/C	30,5	2981	61,5	95	0,01	0,94	0,02	0,03	C	C
32	TPSNEI2L	A/G	1626	64	3165,5	19	0,33	0,01	0,65	0,00	A/G	A/G
33	TPSNEI2253L	C/T	36	837,5	138	19,5	0,03	0,81	0,13	0,02	C	C
34	TPSNEI2362L	C/T	79	3865	136	64,5	0,02	0,93	0,03	0,02	C	C
35	TPSNE4AU	C/T	53	1529	327	58,5	0,03	0,78	0,17	0,03	C	C
36	TPSNE4BL	C/T	414	496	790,5	82	0,23	0,28	0,44	0,05	G	C
37	TPSNE4DL	G/C	97,5	2169,5	93	749	0,03	0,70	0,03	0,24	C	C
38	TPSNE6U	A/G	68,5	47,5	1127,5	107,5	0,05	0,04	0,83	0,08	G	G
39	TPSNE6637U	C/T	28	2168,5	1327	16,5	0,01	0,61	0,37	0,00	C/G	C
40	TPSNE6650L	A/G	169,5	96	3327,5	689	0,04	0,02	0,78	0,16	G	G
41	LMP7E1B595U	A/G	121	62,5	3954,5	68	0,03	0,01	0,94	0,02	G	G
42	T1E41064L	C/T	282	147,5	190	4091	0,06	0,03	0,04	0,87	T	T
43	T2E111383U	C/T	560	2076	287	72	0,19	0,69	0,10	0,02	C	C
44	TPSNE4CU	G/C	484,5	112	1664,5	71	0,21	0,05	0,71	0,03	G	G
45	T1E101030U	A/G	105	69	2809	36,5	0,03	0,02	0,93	0,01	G	G

Table 1: Experimental results for the 45-plex Human Major Histocompatibility Complex. Column 3 shows potential alleles as reported in the literature, columns 4-7 absolute fluorescence levels after background subtraction, columns 8-11 relative fluorescence levels for each of the 4 bases. Differences from 100 % are due to rounding. A base was called whenever its % fluorescence exceeded 30%; Base IDs were confirmed independently by direct sequencing.

seconds.

The reactions were then incubated with 2 μ l of microsphere mix that contained 45 microspheres conjugated to 45 antitags that were complementary to the tags associated with each of the 45 minisequencing primers. The incubation allowed for the hybridization of tag-antitag pairs, and the capture (and thus demultiplexing) of the

minisequencing primers that had now been extended by a single biotinylated ddNTP, onto microspheres. This hybridization was performed in a binding buffer that contained 100 mM Tris-HCL, 1 mM EDTA, and 800 mM NaCl. The hybridization cycle consisted of an initial increase in temperature to 80°C, to allow for the denaturation of all DNA single strand interactions,

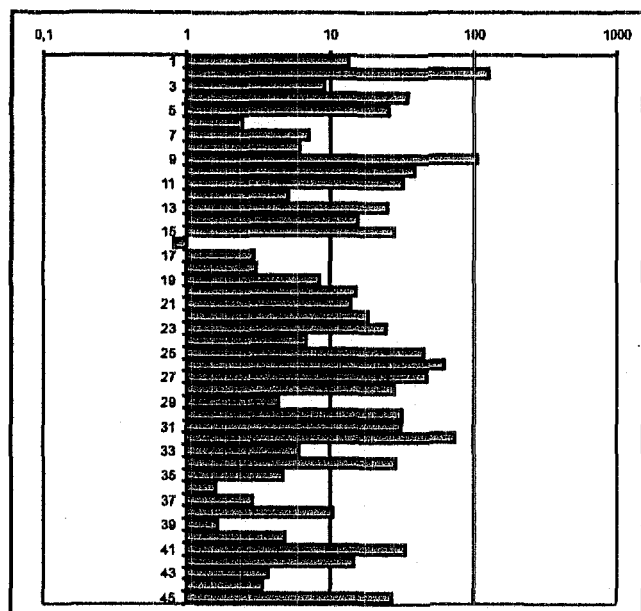


Fig. 5. Signal to noise ratio for 45-plex minisequencing assay. The plot shows the ratio of the absolute signal of the true base to the highest false base on a logarithmic scale.

followed by a stepwise decrease in temperature to 25°C, by holding at 70, 60, 50, 40 and 35°C for 1 minute. This decrease in temperature allowed for the gradual annealing of the specific tag with its complementary antitag on the microsphere. Following hybridization, the microspheres were washed two times with the same buffer, that also contained 0.02% Tween 20, to prevent the microspheres from sticking to each other, as well as the tubes they were contained in. The wash step was performed to remove all excess, unextended biotinylated ddNTPs, that would bind non-specifically to the fluorescent stain. The hybridized minisequencing primers were then resuspended in 35 μ l of buffer that contained Streptavidin conjugated Phycoerythrin (23 μ M, red fluorescent dye), and incubated for 15 minutes at room temperature. The biotin-ddNTP extended minisequencing primers are thus stained with the fluorescent dye, which will be detected on the LUMINEX Flow-Cytometer.

All reactions were transferred to 96 well plates to enable analysis by the LUMINEX Flow-Cytometer. Data was collected and analysed using Microsoft Excel. Table 1 summarizes the data. The results are graphically presented as shown in figure 4. Bases were called whenever the percentage fluorescence for the base was over 30 % of total fluorescence of all four bases. Base IDs for each of the 45 SNP sites were confirmed by direct sequencing of the same PCR amplicons; results were confirmed in all cases except for SNP 36, where multiplexed minisequenc-

ing calls Guanine with 44 % of total fluorescence, while conventional sequencing shows Cytosine. It is not clear if this is a sequencing or a minisequencing problem. The signals for Adenine and Cytosine for this polymorphism are also very strong in the minisequencing assay.

Two additional cases show high Guanine background (Beads 16 and 39). A control experiment with no template sequence (primer-tag duplexes only in the minisequencing reaction) has been conducted, showing high G background for the corresponding beads as well. Figure 5 shows the ratio of the absolute fluorescence value of the true base to the highest false signal in the assay. Note that the PCR product evaporated in the reaction for SNPs 6-9, and there is thus no high absolute fluorescence present.

DISCUSSION

The results presented show that minisequencing can be adapted to be a tool for high-throughput genotyping. The data presented on the 45-plex experiment shows the feasibility of the approach, with only minor problems with high Guanine background in some cases. One of the reasons for this background might be that the assay has been conducted with equal amounts of the four ddNTPs; additional experiments conducted show that modifying the ratio of the ddNTPs influences the background distribution (data not shown). Also, leaving the reaction over night before washing and staining worsens the background problem, high G background being affected the worst. The reason for this is not clear, but by using unequal amounts of the four ddNTPs, the background can be reduced, and we are currently in the process of optimizing the protocols.

High throughput and low cost assays are not feasible without high levels of multiplexing; currently available genotyping tools need to be highly parallelized to satisfy the requirements of pharmacogenomics and of routine SNP analysis in medical institutions. Such applications might involve the scoring of millions of SNPs per day.

Such multiplexing is not possible without a way to "demultiplex" the experiment. We have demonstrated that this can be done using a tag-antitag system, which sorts the signals from the minisequencing reaction to the corresponding beads. The higher the level of multiplexing desired, the more complex becomes the problem of designing such a crosshybridization-free tag-antitag code, and the more relevant becomes a profound thermodynamic algorithm to predict interactions. We have generated a code of 250 such tags using the SBEprimer software and demonstrated its quality.

The selection of appropriate primers for the multiplexing assay is a second crucial requirement. Such primers must not false prime, and they must work together in the same assay. The manual design of minisequencing primers quickly becomes impossible if more than just a

few primers are pooled together. The SBEprimer software can automatically design appropriate primer sets, and thus enables high levels of multiplexing.

One important feature of the SNP assay presented is its applicability to Heterozygote detection. If two different alleles are present, the experiment will show high signal for both bases, and make two base-calls. This is a must for any useful genotyping technology, and easily done with the assay presented here.

The quality of the results obtained on the 45-plex experiment presented with its high ratio of true to false signal indicates that much higher levels of parallel genotyping can be achieved using the flow cytometry based technology with the SBEprimer software package; a 65 plex is currently in preparation, and it is likely that much higher levels of multiplexing are possible.

CONCLUSION

Our results show that minisequencing can be adapted to be a powerful tool for high-throughput, cost-efficient genotyping. The simultaneous screening of 45 polymorphic sites has been demonstrated, with basecalls confirmed by independent sequencing. The manual design of such multiplexed genotyping assays is a laborious process, but can be highly automated using the SBEprimer package presented in this work.

The thermodynamic alignment algorithm used in the SBEprimer program calculates very accurate interaction profiles for oligonucleotides, and the experiments show how careful design of an assay with computer support enabled complex reactions to be carried out with the desired results. We intend to use a similar method to design primer pairs for multiplexed polymerase chain reactions. Clearly, this is presently the bottleneck of all genotyping methods, and a higher level of PCR multiplexing would be highly desirable. We are confident that such experiments will benefit from more accurately determined interactions and automated multiplexing primer design software. SBEprimer can be adapted for such purposes.

REFERENCES

- Allawi, H. and J. SantaLucia (1997). Thermodynamics and nmr of internal g-t mismatches in dna. *Biochemistry* 36, 10581-10594.
- Allawi, H. and J. SantaLucia (1998a). Nearest neighbor parameters for internal g-a mismatches in dna. *Biochemistry* 37, 2170-2179.
- Allawi, H. and J. SantaLucia (1998b). Nearest neighbor parameters of internal a-c mismatches in dna: Sequence dependence and ph effects. *Biochemistry* 37, 9435-9444.
- Allawi, H. and J. SantaLucia (1998c). Thermodynamics and nmr of internal c-t mismatches in dna. *Nucleic Acids Research* 26, 2694-2701.
- Ben-Dor, A., R. Karp, B. Schwikowski, and Z. Yakhini (2000). Universal dna tag systems: A combinatorial design scheme. *J. Comput. Biol.* 7, 503-519.
- Brenner, S. (1997). Methods for sorting polynucleotides using oligonucleotide tags. US Patent 5,604,097.
- Breslauer, K., R. Frank, H. Blöcker, and L. Marky (1986). Predicting dna duplex stability from the base sequence. *Proc. Natl. Acad. Sci. USA* 83, 3746-3750.
- Cai, H., P. White, D. Torney, A. Deshpande, Z. Wang, B. Marrone, and J. Nolan (2000). Flow cytometry-based minisequencing: A new platform for high-throughput single-nucleotide polymorphism scoring. *Genomics* 66, 135-143.
- Cooper, D., B. Smith, H. Cooke, S. Niemann, and J. Schmidtke (1985). An estimate of unique dna sequence heterozygosity in the human genome. *Hum. Genet.* 69, 201-205.
- Dopazo, J. and F. Sobrino (1993). A computer program for the design of pcr primers for diagnosis of highly variable genomes. *Journal of Virological Methods* 41, 157-166.
- Frutos, A., Q. Liu, A. Thiel, A. Sanner, A. Condon, L. Smith, and R. Corn (1997). Demonstration of a word design strategy for dna computing on surfaces. *Nucleic Acids Research* 25, 4748-4757.
- Garey, M. R. and D. S. Johnson (1979). *Computers and Intractability*. New York: Freeman.
- Giegerich, R., F. Meyer, and C. Schleiermacher (1996). Genefisher - software support for the detection of postulated genes. In *Proceedings of the fourth international conference on intelligent systems for molecular biology*. AAAI Press.
- Gotoh, O. and Y. Tagashira (1981). Stabilities of nearest-neighbor doublets in double-helical dna determined by fitting calculated melting profiles to observed profiles. *Biopolymers* 20, 1033-1042.
- Hirschhorn, J. N., P. Sklar, K. Lindblad-Toh, Y.-M. Lim, M. Ruiz-Gutierrez, S. Bolk, B. Langhorst, S. Schaffner, E. Winchester, and E. S. Lander (2000). Sbe-tags: An array-based method for efficient single-nucleotide polymorphism genotyping. *PNAS* 97, 12164-12169.
- Ke, S. and R. Wartell (1995). Influence of neighboring base pairs on the stability of single base bulges and base pairs in a dna fragment. *Biochemistry* 34, 4593-4600.
- LeBlanc, D. and K. Morden (1991). Thermodynamic characterization of deoxyribooligonucleotide duplexes containing bulges. *Biochemistry* 30, 4042-4047.
- Lucas, K., M. Busch, S. Mössinger, and J. Thompson (1991). An improved microcomputer program for finding gene- or family-specific oligonucleotides suitable as primers for polymerase chain reactions or as probes. *Cabios Communication* 7, 525-529.
- Morris, M., D. Shoemaker, R. Davis, and M. Mittmann (1997).

- Methods and compositions for selecting tag nucleic acids and probe arrays. European Patent application, 97302313.
- Nolan, J. and L. Sklar (1998). The emergence of flow cytometry for sensitive, real-time analysis of molecular interactions. *Nat. Biotech.* 16, 633-638.
- Owczarzy, R., P. Vallone, F. Gallo, T. Paner, M. Lane, and A. Benight (1997). Predicting sequence-dependent melting stability of short duplex dna oligomers. *Biopolymers* 44, 217-239.
- Peyret, N., P. Seneviratne, H. Allawi, and J. SantaLucia (1999). Nearest-neighbor thermodynamics and nmr of dna sequences with internal a-a, c-c, g-g and t-t mismatches. *Biochemistry* 38, 3468-3477.
- Quartin, R. and J. Wetmur (1989). Effect of ionic strength on the hybridization of oligonucleotides with reduced charge due to methylphosphonate linkages to unmodified oligodeoxynucleotides containing the complementary sequence. *Biochemistry* 28, 1040-1047.
- Rozen, S. and R. Skaletsky (1998). Primer3. Code available at http://www-genome.wi.mit.edu/genome_software/other/primer3.html.
- Rychlik, W. and R. Rhoads (1989). A computer program for choosing optimal oligonucleotides for filter hybridization, sequencing and in vitro amplification of dna. *Nucleic Acids Research* 17, 8543-8551.
- SantaLucia, J. (1998). A unified view of polymer, dumbbell, and oligonucleotide dna nearest-neighbor thermodynamics. *Proc. Natl. Acad. Sci. USA* 95, 1460-1465.
- SantaLucia Jr., J., H. Allawi, and P. Seneviratne (1996). Improved nearest-neighbor parameters for predicting dna duplex stability. *Biochemistry* 35, 3555-3562.
- Schafer, A. and J. Hawkins (1998). Dna variation and the future of human genetics. *Nature Biotech.* 16, 33-39.
- Smith, T. and M. Waterman (1981). Identification of common molecular subsequences. *J.Mol.Biol.* 147, 195-197.
- Sommer, R. and D. Tautz (1989). Minimal homology requirements for pcr primers. *Nucleic Acids Research* 17, 6749.
- Sugimoto, N., S. Nakano, M. Yoneyama, and K. Honda (1996). Improved thermodynamic parameters and helix initiation factor to predict stability of dna duplexes. *Nucleic Acids Research* 24, 4501-4505.
- Syvänen, A.-C. (1999). From gels to chips: "minisequencing" primer extension for analysis of point mutations and single nucleotide polymorphisms. *Human Mutation* 13, 1-10.
- Turner, D. (1992). Bulges in nucleic acids. *Current Opinion in Structural Biology* 2, 334-337.
- Venter, J., M. Adams, E. Myers, P. Li, et al. (2001). The sequence of the human genome. *Science* 291, 1304-1351.
- White, P. and D. Torney (2001). A "solution" to solid surface measurements. NFCR Newsletter, May 2001, <http://lsdiv.lanl.gov/NFCR/newsletter-My01/may01.html>.